

Adaptive PSO-Based Ensemble Optimization for Histology Image Classification

¹Eid Alkhalidi, ²Ezzatollah Salari

^{1,2}Dept. of Electrical Engineering and Computer Science, University of Toledo, Ohio, United States

Abstract

Computer-aided breast cancer diagnostic systems' credibility relies profoundly upon the accuracy of the models correlating the various enigmatic descriptors to the correct class labels. Consequently, breast H&E samples' classification is one of the most crucial problems of computer vision in the medical field. As an outcome of the accelerated evolution in computational resources, Convolutional neural networks emerged to elicit more intricate features. Nonetheless, CNNs are data greedy and inclined to overfit in the medical field due to the deficient supply of labeled patches. Although Transfer Learning assists in reducing the massive training-data requirement of CNNs, it fails to separate domain-specific layers from the general ones, which often leads to the worst accuracy on unseen samples. The shortcoming mentioned earlier is often attributable to the lack of interpretability of the features extracted at each layer. Moreover, the high number of hyper-parameters associated with ensembles of pre-trained models greatly magnifies the search space, which substantially increases the training time. To overcome these problems, we propose an adaptive particle swarm optimization hyper-parameter selection method that focuses on deriving an ensemble rule for a fixed length of the best-trained models. We first fine-tune a set of pre-trained models on low-resolution images and determine the best combination of them based on the variance of classification errors amongst them rather than their validation loss. Then, we use a PSO whose learning rate is adjusted with a Mamdani fuzzy inference system to infer the ensemble policy. Finally, we train the ensemble thoroughly as a noisy-student model using hyper-parameters obtained from the two search sessions with high-resolution and weakly classified images. The outcomes of our method were compared to different state-of-the-art techniques implemented on the ICIAR 2018 dataset. The proposed approach substantially outperformed previously published procedures.

Keywords

Particle Swarm Optimization, parameters selection, Ensemble, CNNs, Histology Image Classification

1. Introduction

The techniques of microscopic classification of H&E specimens of breast tissue aim to extract distinctive intensity-based determinants that make the cancerous classes more distinguishable [1]. The abnormality detection task is driven by the spatial and spectral information content of the images. As a consequence of the intricate landscape of this problem and its clinical significance, carcinomic stage identification is vital to the medical image processing area [2-6]. The hand-operated examination of the histopathology images demands an immense volume of labor and time of highly educated specialists who are limited and costly [3-5]. Moreover, professionals often find the analysis of histology images challenging and may disagree on the cancerous level of a slide due to individual biases [4, 7]. The medical treatment and the future forecast of the malignancy behavior is reliant upon early

diagnosis [4, 7, 8-9]. The reasons above motivated researchers to exploit to a subset of Artificial Intelligent designs named computer-aided diagnostic systems (CAD) [8].

Various studies verified that Deep Learning (DL) in histopathology had a competing performance to the domain experts [10-11]. The success of DL is credited to its supremacy in deriving features that are deeply embedded into the image, which significantly grows the malignancy types boundary. Besides, the extensive growth in Deep Learning and computational resources, such as GPUs, allowed researchers to devise profoundly sound predictive models for histopathology patches [12, 10]. Several DL schemes were adopted toward medical image classification. One of the leading DL frameworks is the end-to-end training of Convolutional Neural Networks (CNNs) [13-14]. End-to-end training of CNNs performed exceptionally well on benchmark image datasets such as MNIST and CIFAR, which combine millions of annotated images. Even though training models end-to-end was remarkably beneficial and more precise in image classification, it is a data-greedy process that demands a massive dataset. The insufficient quantity of marked samples will prompt the model to learn noise as the number of epochs increases. Multiple studies in the histological images classification revealed that training a DL model using an insufficient number of classified examples is prone to overfitting regardless of how the weights were initialized. The earlier asserted point is the central hindrance of the state-of-art end-to-end procedures, which negatively impairs its performance [5].

To overcome the deficiencies of the end-to-end CNNs training, Transfer Learning arose as a more hybrid alternative. Transfer learning is the reusing of a model trained on a large dataset for a different task [15-16]. Several studies revealed a high correspondence between the accuracy of models trained on benchmark datasets and their performance on other domains [17]. As an example, models that scored well on CIFAR will function well on medical datasets. This high correlation permits industries to incorporate the models that they previously trained into their pipelines for additional predictive tasks. The leading cause of this high correlation is the fact that most of the initial layers of the CNNs learn similar fundamental features that are universal to all images, such as edges and basic shapes. As a consequence, researchers adopted these pre-trained models as feature extractors that were utilized for training less intricate models such as Support Vector Machines (SVMs) [17]. Despite its early success, feature engineering is very reliant on the knowledge of the data scientist and needs a substantial manual tuning of parameters based entirely on intuition [18-19].

A Different approach to transfer learning is fine-tuning. Fine tuning a pre-trained model with the right parameters is a very time-efficient and had better performance than feature engineering [20-21]. However, it is not efficient in terms of the number of parameters [20]. The fact, as mentioned earlier, could result in

overfitting, mainly when training the general layers [1].

Multiple methods consolidated preprocessing techniques such as thresholding, entropy maximization, fuzzy entropy, and sharpening to enhance the quality of segmentation and classification [8, 22-23]. Even though the methods stated earlier, marginally raised the labeling accuracy, they hold fewer peculiarities than what the classification task demands, which oversimplifies the feature space. Considering the fewer parameters, they furthermore tend to underfit significantly with small datasets. However, these methods are usually incorporated as a preprocessing step for a more complicated algorithm.

Analyzing the last layers of the pre-trained models is inefficient in terms of time since the number of trainable parameters change when fine-tuning. Therefore, ensembling a number of models proved to be very effective and more robust alternative [24-27]. Ensembles are a vertical or cascaded combination of the features or the predictions of accurate models to produce a better performing model [1, 24, 26, 28]. For ensembles to perform better than its individual model, the models have to be precise and heterogeneous [1, 24]. The heterogeneity condition ensures that the fused models don't make the same misclassifications, which implies that they learned different features [5].

An active research field in Deep Learning has been dedicated to study the diversification of ensembles by using advanced optimization techniques such as Genetic Algorithm [1]. Due to the time-consuming and the high computational power of ensemble methods, they are only viable for applications similar to medical image classification where time is not a major issue [27, 29-32].

The volume of the different possible configurations of a particular ensemble model is massive which greatly complexifies the search of optimal settings [27, 32]. Due to the impracticality of evaluating every possible ensemble configuration, multiple optimization methods were proposed. Early methods for ensemble optimization were reliant on randomly evaluating a different set of configurations. The random search and the grid search algorithms were common examples of these approaches [29-30]. Other methods relied on the expert knowledge to manually select the hyper-parameters of the system [30, 33]. These methods require extensive training and are not guaranteed to converge [30]. Furthermore, they lack the ability to intelligently evolve based on past trials.

Bio-inspired Computational algorithms proved to be very efficient in solving optimization problems [34]. Particle Swarm Optimization (PSO) is one of the most popular techniques used in NP hard problems and optimization methods for non-convex search-space [35, 36]. The simplicity and ease of implementation of PSO promoted its use across various domains [35, 37-38]. However, the performance of PSO is heavily dependant on the balance between exploration and exploitation by properly setting the right parameters [39].

In this paper we propose a generic method to incorporate the Particle Swarm Optimization to speed up the hyper-parameter selection process of the optimum ensemble of the best pre-trained CNN models. The proposed methodology enable us to have more control over the training process and minimize errors especially between classes that are less distinguishable. Our method selects models based on features that are more sophisticated than the validation

loss of individual models. The proposed method automate the dynamic adjustment of the PSO learning rate by implementing a Mamdani Fuzzy Inference System (FIS). Furthermore, our method address the limitation of labeled images and utilizes the noisy student model approach on the weakly labeled data. Even though this method outperformed many of the state-of-the-art techniques, it is very time-consuming. To the best of our knowledge, this study is one of the first implementations of PSO to learn ensemble hyper-parameters for small histological images dataset.

The remaining parts of this paper are formed in the following order. Section II describes the proposed method in detail. A thorough overview of the methodology is presented in subsection A. The subsections B and C exhibit the preliminaries of PSO and the Mamdani FIS, respectively. Subsection D explains the adaptive PSO using the Mamdani FIS. Section III illustrates the experimental setup, dataset, and implementation details. Section IV offers an elaborate discussion of the experimental results. The conclusion and future research recommendations are offered in Section V.

II. Proposed Method

A. Overview

This section illustrates ensemble optimization, the overall protocol of the proposed method, PSO theory, the Mamdani Fuzzy Inference System (FIS) and the adaptive PSO. The main objective of Ensemble optimization is finding a robust composite of several classifiers in order to cultivate a single larger and more accurate model [1, 26, 36, 40]. Classical ensembling methods such as majority voting lack assumes the optimum weights of the classifiers to be uniform. This assumption undermines the complexity of medical applications. In cancer detection problems, the tissues are expected to be normal unless there is a region that could be classified otherwise. This observation lead to the conclusion that, even weighting of classifiers could results in producing an ensemble with skewed classification towards abnormal labels.

Our proposed method is composed of two main phases. The first phase is centered around training individual models using an adaptive learning rate scheduler to fine-tune networks pre-trained on ImageNet. The second phase is training a linear combination function that combines the output of a fixed number of the best performing networks.

The preprocessing of images is an extremely crucial part of the training process [19, 41-43]. Image augmentation is especially essential to increase the size of the datasets and consequently to increase the generalization of the model [44]. One of the most widely used techniques to enlarge small datasets and increase the robustness of the Artificial Intelligent systems is to generate more data through data augmentation [14, 42, 44-49]. Many data augmentation techniques were proposed, however; different data augmentation strategies with different parameters is known to give varying performance on different domains. Therefore; it is important to search for the augmentation strategy that would yield the optimum performance. However; searching for the parameters that give the optimum performance can add more complexity to the problem. Therefore, we decided to use the Rand-Augment method which eliminate the need to search for optimum parameters for the data augmentation strategy as explained in Sec III (B) [44].

Another important aspect of the training is the proper choice of the optimizer. AdaMax is a stochastic gradient descent algorithm that is used to optimize the categorical cross-entropy loss function. The categorical cross-entropy is defined as follows [12, 50].

$$Loss = - \sum_{\forall x \in set} \sum_{i=1}^4 w_i \cdot \delta_{x,i} \cdot \log(\hat{p}_x, i) \quad (1)$$

Where set is the set of all training dataset images and i refers to the class index whose maximum value is four classes in this problem, \hat{p}_x is the model prediction of image x belonging to class i , and $\delta_{x,i}$ equals 1 when image x belong to class i and zero otherwise.

The performance of AdaMax was compared on the inception model with and without an adaptive learning rate scheduler [51]. Figure 1 shows the training of InceptionV3 model without adaptive learning rate scheduler, while figure 2 shows the training history of the same network and optimizer with an adaptive learning rate scheduler. As the figure 1 and 2 show, the AdaMax proved to be more robust and with a steadily decreasing validation loss curve when the learning rate gets adjusted with accordance to how well it performed in previous epochs. The AdaMax relies on initializing the bias correction based on the moving average [50].

CNNs have a massive search space of hyper-parameters that need to be tuned. The search of the most optimum set of hyper-parameters is often time-consuming [32]. Another problem is that using high resolution images during training will reduce the maximum batchsize allowed due to the limited memory available which reduces the generalization. In order to solve these problems, we used a lower-dimensional representation of the datasets during the first phase of the training in order to identify the most auspicious hyper-parameters.

Furthermore; using a lower data representation during training than doesn't only reduce the search space complexity, but is also known to reduce the disparateness of the dimensions of distinct image components between training and testing [48, 52]. The details of the used resolutions are presented in 3.2. Besides, previous research was conducted to find the optimum freeze-layer. We found that the domain-specific layers typically include the last two fifths of the total number of the trainable parameters [1]. Consequently, we fixed the freeze-layer as well as the length of the ensemble in this work due to time cost of their optimization.

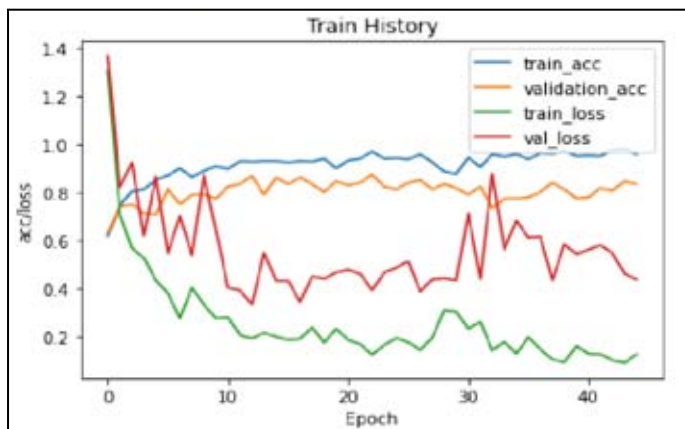


Fig. 1: The training history of the Pretrained InceptionV3 with ADAMAX and without learning rate adaptation.

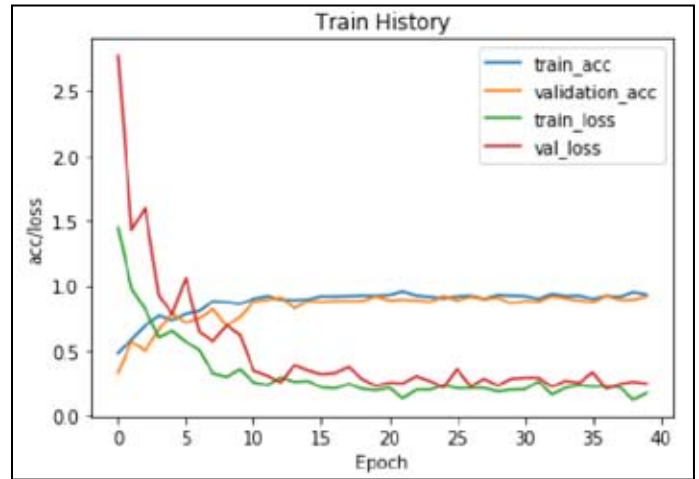


Fig. 2: The training history of the Pretrained InceptionV3 with learning rate adaptation.

B. Particle Swarm Optimization

Extensive research was conducted to study the animals swarm societal behavior [28, 53, 54, 36, 22]. Swarm-based optimization techniques were developed to mimic the social behaviour of groups of animals that are navigating their way cooperatively to find food resources for the sake of evolving optimum solution for NP-hard problems [39, 55, 28]. PSO is a biologically inspired heuristic stochastic optimization algorithm that aims at finding the global minimum of objective functions with complex search space [38, 53, 54, 28, 3, 56, 57]. It was first proposed by Eberhart and Kennedy in 1995 as a paradigm for Neural Networks weights optimization [58]. Contrary to prior population-based approaches, PSO was motivated by the apprehension that the interaction of birds ameliorate their hunt for resources [58].

Evolutionary computational algorithms showed superior performance in optimization problems both in terms of time-efficiency and loss minimization [58]. PSO in particular is one of the most effective of these methods [58, 53, 39, 56]. For the those reasons, PSO has been successfully applied in numerous domains.

Analogous to other evolutionary-based algorithms, PSO starts with initialization of a generation that has a specified number of candidate solutions called particles. The evolution of each generation of particles represent the pursuit of survival by a flock of birds. Every particle is a vector that depicts a likely solution. In each epoch, particles move in the defined finite search space of the problem based on how well the whole swarm of birds is performing. The swarm share the best position achieved by the whole swarm, which is the global best position and fitness. The velocity and the position of the particle are each updated based on the equation 2 and 3.

$$V_{ij}^{t+1} = wV_{ij}^t + c_1r_1^t (pbest_{ij} - X_{ij}^t) + c_2r_2^t (gbest_j - X_{ij}^t) \quad (2)$$

$$x^{t+1} = x^t + V^{t+1} \quad (3)$$

where V_{ij}^{t+1} is the updated velocity, V^t is the current velocity, w is the inertia weight, $c1$ and $c2$ are acceleration constants referred to as the cognitive learning rate, $gbest$ and $pbest$ are the global and the population best positions respectively, and $r1, r2$ are random numbers evenly distributed between $[0,1]$ [3].

PSO has many advantages compared to other optimization methods. Some of these advantages of PSO are its robustness, ease of implementation and remarkable efficiency in reaching accurate approximations [28]. PSO was used to find the best linear combination weights of the horizontally stacked fine-tuned models.

The convergence of PSO is highly reliant on the proper tuning of its learning hyperparameters. In particular, PSO is highly sensitive to the Inertia Weight w , the accelerate constants $c1$ and $c2$, the number of particles, the number of generations and the maximum velocity [3, 39]. These parameters have to be chosen delicately in order to avoid divergence or premature convergence. The inertia weight variable w is the most important parameter to tune in order to avoid getting trapped in a local minima. It regulates degree of the influence that the earlier velocity has on the updated velocity [39]. Therefore, it controls the speed at which a particle moves in the search space. The larger the inertia weight is, the faster the particle moves and the less likely it will be to get trapped on a local minimum. However, this could also mean that the PSO can't find the global minimum if the w is too large. In contrast, if w is too small, it is very likely for the PSO to converge prematurely. This phenomenon is recognized in the literature as the exploitation and the exploration trade-off [39]. The ability of the particles to ameliorate the local solution is referred to as exploitation while exploration refers to their ability to escape local minima [39].

The Inertia weight is not the only parameter that is accountable for governing the exploration during the PSO evolution. The accelerate parameters $c1$ and $c2$ also play a major role on managing the swiftness of the particles in the search space. The $c1$, and the $c2$ are often referred to as the cognitive and the social constants in the PSO literature [59, 58]. The cognitive constant $c1$ controls the quantum of acceleration upon which the particle's velocity updated towards its personal best value. When $c1 = 0$, the particle's is not aware of its personal best as can be seen in equation 2. It is evident that when both of the constants $c1, c2$ are set to zero, the particles has no knowledge of either the p_{best} and the g_{best} which decrease the likelihood to find the global minimum. Similarly, $c2$ is named the social constant because it regulates the acceleration of the particle towards the global best [3]. The accelerate parameters were considerably researched in order to find a comprehensive balance between them. However, different applications yielded different results. In general it was proven in previous studies that the variation of the accelerate constants did not show improvements to the performance of the PSO [3].

The number of swarms particles was not adapted with the FIS since it does not have a significant impact on the performance of the PSO when the size of population is above 50 [60, 3].

The tuning of V_{max} is another problem that deals with the exploration vs exploitation trade-off. As an illustration, when the v_{max} is too large the exploration capability is high and the particles might move away from the optimum region and vice versa. Section II (C) explains PSO parameters adaptation with more emphasis on the impact of the variation of the inertia weight on the momentum of the training.

PSO has many advantages compared to other optimization methods. Some of these advantages of PSO are its robustness, ease of implementation and remarkable efficiency in reaching

accurate approximations [28]. PSO was used to find the best linear combination weights of the horizontally stacked fine-tuned models.

C. Mamdani FIS

Tuning the learning parameters has been a critical issue for the performance of PSO. Several methods were proposed to adaptively control specific parameters optimization algorithms based on some inputs. One of the most important methods to control the performance of evolutionary-based algorithms during the training process is the knowledge-based Fuzzy Inference System [55].

Unlike binary logic that assumes that a particular object either belongs or does not belong to a specific class, the fuzzy sets theory was introduced by Lotfi Zadah in the 1960s to describe things that belong to a multiple fuzzy sets with a varying degree of belongingness [61, 34, 55, 62, 63, 64]. While probabilistic inference methods dealt with the randomness aspect of a particular system, Fuzzy Inference Systems (FIS) were designed to map fuzzy inputs that belong to fuzzy sets with varying degrees to fuzzy outputs. Fuzzy inference deals with parameters that belong to classes that are not clearly defined, while probability deals with the uncertainty of association. FIS has been extensively applied to the adaptation of parameters for numerous algorithms and has succeeded in incorporating the expert knowledge into the optimization process for applications where the rules are vague or hard to express mathematically.

A fuzzy set is a set whose elements have a degree of association to one or more fuzzy classes [65, 66, 62].

Hence, fuzzy inference refers to the steps of the procedure required to associate elements of fuzzy inputs to one or more fuzzy outputs [62].

The first step for the Mamdani FIS is the fuzzification of the inputs and the outputs. Fuzzification is the process of converting the deterministic inputs into fuzzy elements [62]. The fuzzification is achieved by defining a membership function that draw soft boundaries between fuzzy classes.

There are different types of membership functions that are used to model the sliding degree of membership of crisp input to all possible fuzzy sets. Some of these membership functions include the triangular, trapezoidal and Gaussian functions [67]. Even though the Gaussian functions function is more accurate in describing the fuzziness of the inputs than triangular and trapezoidal, it is less time-efficient [67]. Equation 4 represent the membership used to fuzzify the inputs.

$$\mu(x, a, b, c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \quad (4)$$

where a and b are the boundaries of the fuzzy set, c is its center and a $a \neq b \neq c$ [67].

The fuzzy rules are formed in an IF-THEN linguistic format based on the experience. The fuzzified inputs are evaluated against a set of antecedent rules. The fuzzy rules are either an intersection or a union of antecedents based on the correlation minimum or the correlation product to calculate the consequents [62]. The resultant consequents are cropped forms of the antecedents.

These consequents are then combined to form a unified aggregate fuzzy output. The last step of the Mamdani FIS is to defuzzification of the aggregate fuzzy set. One of the most methods of defuzzification is to calculate the center of gravity shown at equation 5 [62].

$$CenterofGravity = \frac{\sum_{x=a}^b \mu_A(x)x}{\sum_{x=a}^b \mu_A(x)} \quad (5)$$

The details of the implementation of the Mamdani FIS for the inertia weight adaptation based on the PSO performance is thoroughly explained in Section D.

As explained on section B, the inertia weight w is the most pivotal parameter to tune in order for the PSO to converge smoothly. Therefore, our approach will emphasize on creating fuzzy rules based on the previous fitnesses values to guide the adaptation of w .

D. Adaptive Particle Swarm Optimization

For practical considerations, the user usually chooses the hyper-parameters of the PSO prior to training. However, setting up the learning rates as constants might not lead to the best performance. Many methods were proposed to control and adapt the hyper-parameters during the evolution of PSO in order to increase efficiency and accuracy without having to repeat the process multiple times. As explained previously, the inertia weight has the most impact on the smoothness of PSO convergence. Our proposed method focuses on adapting the inertia weight by monitoring the previous fitness of the PSO as well as the normalized distances between the current particle and its personal and global best positions. Our method was inspired by Zamli’s method to control the velocity by monitoring the PSO parameters [39]. The Mamdani FIS has shown a great success in previous studies in controlling the PSO parameters, thus it is wise to think it will work in this problem as well. This paper demonstrate the effectiveness of the Mamdani FIS in controlling the PSO for finding the most optimum ensemble rule for histopathology image classification.

The Mamdani FIS is one of the most effective in detecting the region that contains the optimum solutions. Ideally, the inertia weight w should be set to its highest value to increase the PSO exploration capacity. Once the region of interest in the search space is reached, the inertia weight should be set to a low value in order to refine the local solution and increase the PSO’s exploitation capacity. If the Mamdani FIS detects the region of the optimum solution, the inertia weight get adapted with accordance to the evaluation metrics defined in equations 7, 8 and 6.

The first metric based on which the performance of PSO is monitored is the Zamli Normalized Current Fitness (NCF) which computes the current fitness of the particle relative to the maximum and the minimum fitnesses as shown in equation 6 [39]. The other two monitoring metrics are the Minkowaski distances of the particle’s current position and the personal best position and the global best position as shown in equation 7 and equation 8 respectively.

$$NCF = \frac{CF - MinF}{MaxF - MinF} \times 100 \quad (6)$$

where CF is the current fitness, MinF is the minimum fitness and maxF is the maximum fitness. The distance metrics were chosen to be the Minkowaski distance since it was proven to be the most accurate distance metric compared to the Manhattan and the Euclidean distances [68]. The Manhattan distance gives more approximate value than the real value while the Euclidean usually gives less value. Several studies were conducted to investigate the optimal values of the where $pbest$ is the personal best position, $gbest$ is the global best position and p is the parameter that specifies the type of the used distance.

$$dp = \frac{[pbest - x]^p}{Dmax} \quad (7)$$

$$dg = \frac{[gbest - x]^p}{Dmax} \quad (8)$$

For the Euclidean distance, $p = 2$ while it is unity for the Manhattan distance. The Minkowski distance uses any value between 1 and 2 for p . The comparison of distance metrics studies recommended $p = 1.54$, however we set $p = 1.5$ which is the middle value between the Manhattan and Euclidean distances for the sake of simplicity [68].

For the optimum region detection, the following Mamdani fuzzy rules were defined as shown in Table 1.

Table 1: The Mamdani Fuzzy Rules

Rule number	antecedents	consequent
1	d1 = high, d2 = high, NCF = high	w = high
2	d1 = low, d2 = medium, NCF = medium	w = high
3	d1 = low, d2 = low, NCF = medium	w = high
4	d1 = low, d2 = low, NCF = low	w = low

When all the metrics are low, the optimal solution is close. Therefore, the value of w is set to low in order to increase the exploitation to refine the local search. Otherwise, the region of interest is not near yet. As a result, the exploration is increased by giving a high value for w . Figure 3 shows the fuzzy membership functions of the antecedents and the consequents anf the FIS.

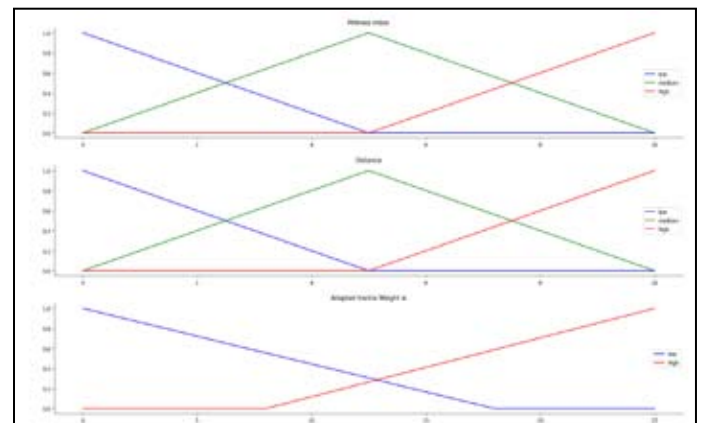


Fig. 3: The Mamdani FIS membership functions for the fuzzification of the PSO Inertia Weight

III. Experimental Setup

A. Datasets

We conducted our experiments using the PyTorch library; which is available on the public domain with NVIDIA Geo-Force RTX 2080 GPU that is connected to an AMD Ryzen Threadripper 1950X 16-Core processor operated with Linux OS. The validation of our technique was performed on a publicly available benchmark dataset as detailed in A.

To empirically assess the performance of our proposed method, we used the ICIAR 2018 dataset. There are various histology images datasets available for the public domain. Nevertheless; the ICIAR is the most realistic datasets due to the fact that it contains a very small number of labeled data which resembles the practical problem encountered in this field.

The ICIAR 2018 datasets contain 400 patches that are labeled into four classes as shown in figure 4. The test data contain 100 unclassified patches that are used by the competition for algorithms evaluation [46].

Algorithm 1 Calculate $y = z^a$

Require: *MaxIter*, *c1*, *c2*, *targetError*, *numParticles*, *numVariables*, *r1* and *r2*

```

Define Mamdani Fuzzy Rules
Initialize the PSO population positions
Initialize best particle positions
Initialize best particle fitness
Initialize best global positions
Initialize best global fitness
Initialize best velocity
while Iter < MaxIter do
  for particle in swarm do
    Calculate fitness
    Calculate Relative Fitness
    Calculate the Minkowski distance index GettheFuzzyBlockactionw
  Update the particle best fitness
  Update the global best fitness
  Update the velocity vector
  Update the position
  Iter += 1
  end for
end while
    
```

The objective of the dataset is to motivate research for the breast cancer detection problem which is one of the leading causes of death amongst women worldwide [46]. ICIAR 2018 also aims at the overcoming the manual analysis of images which requires very specialized knowledge and often lead to non-consensual diagnoses [69]. Examples of the labeled patches are shown in figure 4.

B. Implementation Details

This section exhibits the implementation details and considerations that were not covered on Section II. One of the implemented techniques that proved to increase the generalization in immense CNNs is the early stopping of training [70]. Furthermore; increasing the momentum to 90% of the batch normalization layers while freezing the general layers improved the speed of convergence of fine-tuning of the individual models [71-72].

For cross-validation the dataset was initially divided into training, validation and testing datasets with 90%, 5% and 5% of the labeled data respectively [73]. Cross-validation was used for training the individual models in the first phase. For the ensemble training with the adaptive PSO, the noisy student self-training method was used [25]. The noisy student semi-supervised method aims at increasing the accuracy of the training by increasing the size of the labeled data by weakly labeling a portion of the actual test

data. The student model; which is the PSO ensemble; is then trained on the whole training data as well as the new weakly noisy labeled data. Adding noise to the labels acquired by the semi-supervised method proved to increase the overall robustness of the model [25]. Affine transformation, rand-Augment and other known augmentation method were implemented to increase the dataset after normalization [44, 45, 74].

The size of the patches is key to the accuracy and the robustness of the CNN models.

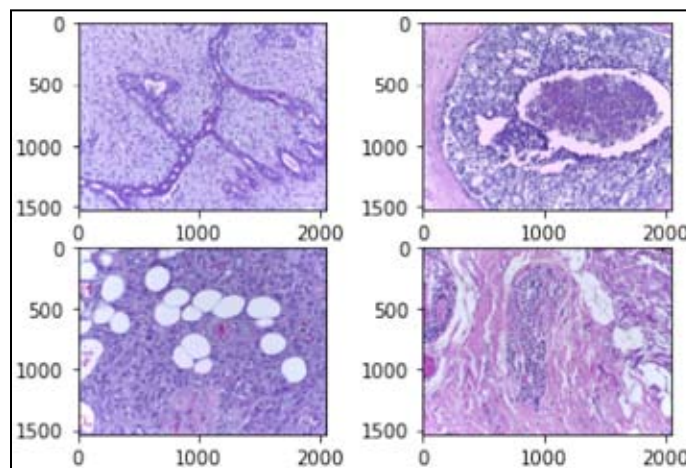


Fig. 4: Benign, InSitu, Invasive and Normal H&E stained breast samples (from top left to bottom right)

However; training the images using a lower dimensional representation than that of the test data proved to increase the robustness of the models [52, 48]. The images size during the training and testing was 600×600 and 800×800 respectively.

IV. Results and Discussion

To empirically assess the performance proposed method, we used various evaluation metrics to compare our approach to other state-of-the-art methods. F1-score, precision, recall, AUC and other widely known measures in the DL field [75, 15]. The results attained by our method are shown in table 2, 3 and 4. The baseline model to which we compare our results was an ensemble with the same fine-tuned models. The results show that optimizing the ensemble with the adaptive PSO improved results significantly compared to the baseline model. The precision, recall and F1-score improved were specifically enhanced in the normal patches. It is evident that the skewed classification due to the high rate of false positives was an obstacle faced by most classifiers. Table 4 show that our approach outperformed the weighted average method by 2% accuracy.

Table 2: Majority Voting Classification Report

{Class. Report}	precision	recall	F1-score
Benign	1.00	0.37	0.54
InSitu	0.60	1.00	0.75
Invasive	1.00	0.93	0.97
Normal	0.97	1.00	0.98
micro avg	0.82	0.82	0.82
macro avg	0.89	0.82	0.81
weighted avg	0.89	0.82	0.81
samples avg	0.82	0.82	0.82

Table 3: Proposed Voting Classification Report

{Class. Report}	precision	recall	F1-score
Benign	1.00	0.73	0.85
In Situ	0.97	1.00	0.98
Invasive	1.00	1.00	1.00
Normal	0.81	1.00	0.90
micro avg	0.93	0.93	0.93
macro avg	0.94	0.93	0.93
weighted avg	0.94	0.93	0.93
samples avg	0.93	0.93	0.93

Table 4: Evaluation of the Proposed Method Compared to Average voting

Eval. metric	Proposed	Avg.
P. AUC ROC	0.9983	0.9934
L. AUC ROC	0.966	0.95
L. accuracy	0.951	0.925
P. precision	0.99	0.988
L. precision	0.9033	0.883
L. log loss	2.108	2.590
L. coverage error	1.157	1.225
P. coverage error	1.109	1.133
L. LRAP	0.963	0.943
P. LRAP	0.959	0.9527
ranking loss	0.052	0.075
ICIAR acc.	89%	87%

V. Conclusion

In this paper, we developed a knowledge-based evolutionary optimization framework for learning the ensemble rule. Our approach aimed at decreasing the number of trials of ensemble optimization with different hyper-parameters by leveraging the fuzzy sets theory. The performance analysis of our approach demonstrated that ensemble optimization of properly fine-tuned models has potentially increased the accuracy of the networks. Even though leveraging the predictions of fine-tuned networks is important to the overall accuracy of the ensemble, it is limited to the degree of error variance amongst the chosen models. Furthermore, the ensemble optimization and the proper choice of the freeze-layer are considered separately which increases the training time significantly. The adaptive variation of the learning rate and the freeze-layer of the networks during fine-tuning could significantly reduce the training time. Our future research will focus on incorporating the knowledge-base Fuzzy Inference Systems into the tuning of the learning rate and the number of trainable weights concurrently.

References

- [1] E. Alkhaldi, "Optimized Heterogeneous Ensemble for Histological Image Classification," Vol. 8, no. 95, pp. 1–2, 2019.
- [2] G. He, L.; Long, LR; Antani, S. and Thoma, "Computer Assisted Diagnosis in Histopathology.," vol. 3, pp. 272–287, 2009.
- [3] Y. He, W. J. Ma, and J. P. Zhang, "The Parameters Selection of PSO Algorithm influencing On performance of Fault Diagnosis," MATEC Web of Conferences, vol. 63, no. 2016, p. 02019, 2016.
- [4] H. M. Ahmad, S. Ghuffar, and K. Khurshid, "Classification of Breast Cancer Histology Images Using Transfer Learning," Proceedings of 2019 16th International Bhurban Conference on Applied Sciences and Technology, IBCAST 2019, pp. 328–332, 2019.
- [5] A. Pimkin, G. Makarchuk, V. Kondratenko, M. Pisov, E. Krivov, and M. Belyaev, "Ensembling Neural Networks for Digital Pathology Images Classification and Segmentation," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10882 LNCS, pp. 877– 886, 2018.
- [6] M. Z. Alom, T. Aspiras, T. M. Taha, V. K. Asari, T. Bowen, D. Billiter, and S. Arkell, "Advanced Deep Convolutional Neural Network Approaches for Digital Pathology Image Analysis: a comprehensive evaluation with different use cases," vol. 2, 2019.
- [7] N. Brancati, G. De Pietro, M. Frucci, and D. Riccio, "A Deep Learning Approach for Breast Invasive Ductal Carcinoma Detection and Lymphoma Multi-Classification in Histological Images," IEEE Access, vol. 7, pp. 44709–44720, 2019.
- [8] T. A. Azevedo Tosta, L. A. Neves, and M. Z. do Nascimento, "Segmentation methods of HE-stained histological images of lymphoma: A review," Informatics in Medicine Unlocked, vol. 9, no. February, pp. 35–43, 2017.
- [9] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks," IEEE Access, vol. 6, pp. 24680–24693, 2018.
- [10] J.-w. Sidhom and A. S. Baras, "Convolving Pre-Trained Convolutional Neural Networks at Various Magnifications to Extract Diagnostic Features for Digital Pathology," bioRxiv Bioinformatics, 2018. 14
- [11] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1313–1321, 2016.
- [12] A. Amidi, S. Amidi, D. Vlachakis, V. Megalooikonomou, N. Paragios, and E. I. Zacharaki, "EnzyNet: Enzyme classification using 3D convolutional neural networks on spatial representation," PeerJ, vol. 2018, no. 5, pp. 1–11, 2018.
- [13] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. DeMarvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation," IEEE Transactions on Medical Imaging, vol. 37, no. 2, pp. 384–395, 2018.
- [14] Y. Shi, T. Qin, Y. Liu, J. Lu, Y. Gao, and D. Shen, "Automatic Data Augmentation by Learning the Deterministic Policy,"

- tech. rep., 2019.
- [15] Y. Takamitsu and Y. Orita, "Effect of glomerular change on the electrolyte reabsorption of the renal tubule in glomerulonephritis (author's transl)," 1978.
- [16] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," 36th International Conference on Machine Learning, ICML 2019, vol. 2019-June, no. 2018, pp. 4815–4826, 2019.
- [17] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2019-June, pp. 2656–2666, 2019.
- [18] Q. Sun, Y. Liu, Z. Chen, T.-S. Chua, and B. Schiele, "Meta-Transfer Learning through Hard Tasks," pp. 1–14, 2019.
- [19] V. R. Elgin Christo, H. Khanna Nehemiah, B. Minu, and A. Kannan, "Correlationbased ensemble feature selection using bioinspired algorithms and classification using backpropagation neural network," Computational and Mathematical Methods in Medicine, vol. 2019, 2019.
- [20] N. Houlsby, A. Giurgiu, S. Jastrzȳbski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," 36th International Conference on Machine Learning, ICML 2019, vol. 2019-June, pp. 4944–4953, 2019.
- [21] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.
- [22] R. Krishna Priya, T. Chelliah, K. Chandrasekaran, and K. Subramanian, "Brain tumor segmentation in MRI images using integrated modified PSO-fuzzy approach," International Arab Journal of Information Technology, vol. 12, no. 6A, pp. 797–804, 2015.
- [23] Öztürk and B. Akdemir, "Effects of Histopathological Image Pre-processing on Convolutional Neural Networks," Procedia Computer Science, vol. 132, no. June, pp. 396–403, 2018.
- [24] T. G. Dietterich, "Ensemble methods in machine learning," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 1857 LNCS, pp. 1–15, 2000.
- [25] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with Noisy Student improves ImageNet classification," 2019.
- [26] J. Xie, B. Xu, and Z. Chuang, "Horizontal and Vertical Ensemble with Deep Representation for Classification," 2013.
- [27] J. C. Lévesque, C. Gagné, and R. Sabourin, "Bayesian hyperparameter optimization for ensemble learning," 32nd Conference on Uncertainty in Artificial Intelligence 2016, UAI 2016, pp. 437–446, 2016.
- [28] F. Ye, Particle swarm optimization-based automatic parameter selection for deep neural networks and its applications in large-scale and high-dimensional data, vol. 12. 2017.
- [29] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," IEEE Access, vol. 7, pp. 180235–180243, 2019.
- [30] M. Claesen and B. De Moor, "Hyperparameter Search in Machine Learning," pp. 10–14, 2015.
- [31] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Long short term memory hyperparameter optimization for a neural network based emotion recognition framework," IEEE Access, vol. 6, pp. 49325–49338, 2018.
- [32] T. Hinz, N. Navarro-Guerrero, S. Magg, and S. Wermter, "Speeding up the Hyperparameter Optimization of Deep Convolutional Neural Networks," International Journal of Computational Intelligence and Applications, vol. 17, no. 2, pp. 1–15, 2018.
- [33] C. L. Rate, "Learning Rate Schedules and Adaptive Learning Rate Methods for Deep Learning Time-Based Decay," pp. 1–10, 2017.
- [34] R. K. Singh, Fundamentals of natural representation, vol. 9. 2018.
- [35] H. J. Escalante, M. Montes, and L. E. Sucar, "Particle swarm model selection," Journal of Machine Learning Research, vol. 10, pp. 405–440, 2009.
- [36] H. J. Escalante, M. Montes, and E. Sucar, "Ensemble particle swarm model selection," Proceedings of the International Joint Conference on Neural Networks, 2010.
- [37] Y. Zhang, X. Xiong, and Q. Zhang, "An improved self-adaptive PSO algorithm with detection function for multimodal function optimization problems," Mathematical Problems in Engineering, vol. 2013, no. iii, 2013.
- [38] S. Darzi, T. S. Kiong, and B. Salem, "Overview of Particle Swarm Optimization (PSO) on its Applications and Methods," Australian Journal of Basic and Applied Sciences, vol. 7, no. 2, pp. 490–499, 2013.
- [39] K. Z. Zamli, B. S. Ahmed, T. Mahmoud, and W. Afzal, "Fuzzy adaptive tuning of a particle swarm optimization algorithm for variable-strength combinatorial test suite generation," Swarm Intelligence - Volume 3: Applications, pp. 639–662, 2018.
- [40] L. Nanni, S. Ghidoni, and S. Brahmam, "Ensemble of convolutional neural networks for bioimage classification," Applied Computing and Informatics, 2020.
- [41] B. Abdikenov, Z. Iklassov, A. Sharipov, S. Hussain, and P. K. Jamwal, "Analytics of Heterogeneous Breast Cancer Data Using Neuroevolution," IEEE Access, vol. 7, pp. 18050–18060, 2019.
- [42] S. Tabik, D. Peralta, A. Herrera-Poyatos, and F. Herrera, "A snapshot of image Pre-Processing for convolutional neural networks: Case study of MNIST," International Journal of Computational Intelligence Systems, vol. 10, no. 1, pp. 555–568, 2017.
- [43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2323, 1998.
- [44] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," 2019.
- [45] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," 2017.
- [46] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. D. Vu, M. N. N. To, E. Kim, J. T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, and P. Aguiar, "BACH: Grand challenge on breast cancer histology images," Medical Image Analysis, vol. 56, pp. 122–139, aug 2019.

- [47] M. Safwan, S. Saketh Chennamsetty, A. Kori, V. Alex Kollerathu, and G. Krishnamurthi, "Classification of Breast Cancer and Grading of Diabetic Retinopathy Macular Edema using Ensemble of Pre-trained Convolutional Neural Networks," no. Midl 2018, pp. 1–14.
- [48] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy: FixEfficientNet," pp. 12–16, 2020.
- [49] K. Lata, M. Dave, and N. K.N., "Data Augmentation Using Generative Adversarial Network," SSRN Electronic Journal, pp. 1–14, 2019.
- [50] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–15, 2015.
- [51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-Decem, pp. 2818–2826, 2016.
- [52] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy," arXiv, jun 2019.
- [53] M. Clerc, "Discrete Particle Swarm Optimization, illustrated by the Traveling Salesman Problem," pp. 219–239, 2004.
- [54] Y. Chunming and D. Simon, "A new particle swarm optimization technique," Proceedings - 18th International Conference on Systems Engineering, IICSEng 2005, vol. 2005, pp. 164–169, 2005.
- [55] M. Abdulgader and D. Kaur, "Evolving Mamdani Fuzzy Rules Using Swarm Algorithms for Accurate Data Classification," IEEE Access, vol. 7, pp. 175907–175916, 2019.
- [56] M. F. Tasgetiren, P. N. Suganthan, and Q. K. Pan, "A discrete particle swarm optimization algorithm for the generalized traveling salesman problem," Proceedings of GECCO 2007: Genetic and Evolutionary Computation Conference, no. May 2014, pp. 158–167, 2007.
- [57] S. Strasser, R. Goodman, J. Sheppard, and S. Butcher, "A new discrete Particle Swarm Optimization algorithm," GECCO 2016 - Proceedings of the 2016 Genetic and Evolutionary Computation Conference, pp. 53–60, 2016.
- [58] R. Eberhart and J. Kennedy, "New optimizer using particle swarm theory," Proceedings of the International Symposium on Micro Machine and Human Science, pp. 39–43, 1995.
- [59] J. Kennedy and R. C. Eberhart, "Discrete binary version of the particle swarm algorithm," Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, vol. 5, pp. 4104–4108, 1997.
- [60] Y. Shi and R. Eberhart, "Modified particle swarm optimizer," in Proceedings of the IEEE Conference on Evolutionary Computation, ICEC, pp. 69–73, IEEE, 1998.
- [61] E. Chan, H. Zhu, and W. Bazzi, "Fuzzy Logic and Probability Theory," pp. 1 — 7, 2000.
- [62] M. Negnevitsky, Artificial Intelligence.
- [63] W. Mei, "Formalization of Fuzzy Control in Possibility Theory via Rule Extraction," IEEE Access, vol. 7, pp. 90115–90124, 2019.
- [64] M. Mazandarani and X. Li, "Fractional Fuzzy Inference System: The New Generation of Fuzzy Inference Systems," IEEE Access, vol. 8, pp. 126066–126082, 2020.
- [65] C.-H. Chiu, "Adaptive Fuzzy Control Strategy for a Single-Wheel Transportation Vehicle," IEEE Access, vol. 7, pp. 113272–113283, 2019.
- [66] B. M. Keneni, D. Kaur, A. Al Bataineh, V. K. Devabhaktuni, A. Y. Javaid, J. D. Zaiantz, and R. P. Marinier, "Evolving Rule-Based Explainable Artificial Intelligence for Unmanned Aerial Vehicles," IEEE Access, vol. 7, pp. 17001–17016, 2019.
- [67] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, "Feature Extraction Foundations and Applications," tech. rep., 2006.
- [68] R. Shahid, S. Bertazzon, M. L. Knudtson, and W. A. Ghali, "Comparison of distance measures in spatial analytical modeling for health service planning," BMC Health Services Research, vol. 9, no. May 2014, 2009.
- [69] H. Cao, S. Bernard, L. Heutte, and R. Sabourin, "Improve the Performance of Transfer Learning Without Fine-Tuning Using Dissimilarity-Based Multi-view Learning for Breast Cancer Histology Images," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10882 LNCS, pp. 779–787, 2018.
- [70] R. Caruana and S. Lawrence, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping," tech. rep.
- [71] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 32nd International Conference on Machine Learning, ICML 2015, vol. 1, pp. 448–456, 2015.
- [72] L. Huang, D. Yang, B. Lang, and J. Deng, "Decorrelated Batch Normalization," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 791–800, 2018.
- [73] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," International Joint Conference of Artificial Intelligence, no. 0, pp. 0–6, 1995.
- [74] K. Kumar, Kuntal, Sudeep, "2016 IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT) : proceedings : 20-21 May 2016, Bengaluru, India," pp. 1778–1781, 2016.
- [75] P. A. Flach, "The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics," Proceedings, Twentieth International Conference on Machine Learning, vol. 1, pp. 194–201, 2003.